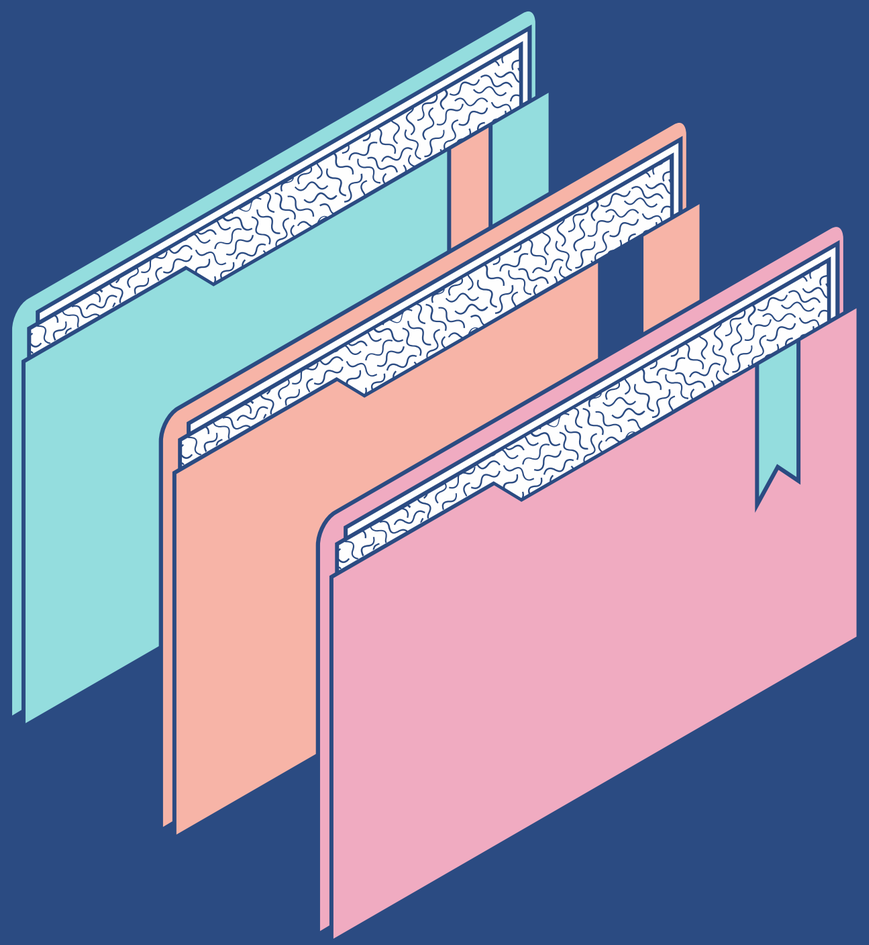
DATA SCIENCE II | CODERHOUSE

Student Performance

Autora: Julieta Milagros Méndez

Comisión: 61755

Año: 2025



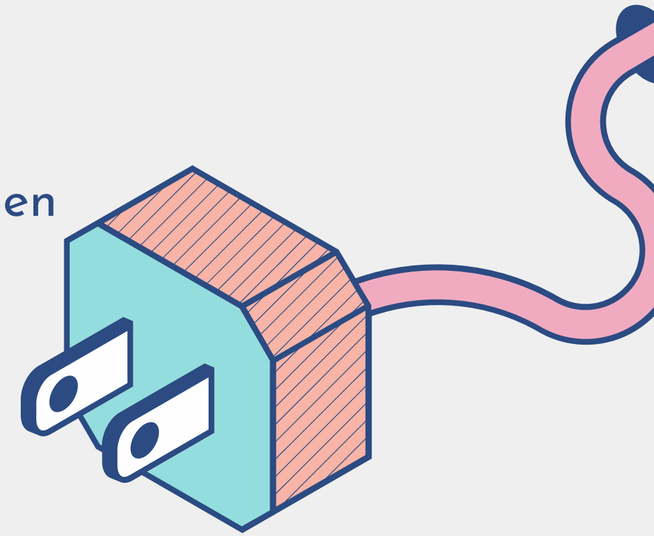
Índice

TÓPICOS DE LA PRESENTACIÓN:

- Introducción
- Análisis descriptivo
- Pre procesamiento de datos
- Entrenamiento de modelos
- Validación cruzada
- Conclusión

Descripción de la temática

Se analizarán los factores que influyen en el rendimiento académico de los estudiantes. En un contexto educativo, los hábitos de estudio, la asistencia a clases, el nivel de involucramiento de los padres y otros aspectos sociales y personales juegan un papel crucial en el desempeño de los estudiantes. Este estudio busca entender cómo estas variables interactúan entre sí y cómo, al predecir el puntaje en los exámenes, se pueden identificar patrones que ayuden a mejorar la calidad educativa.



https://www.kaggle.com/datasets/lainguyn123/student-performance-factors

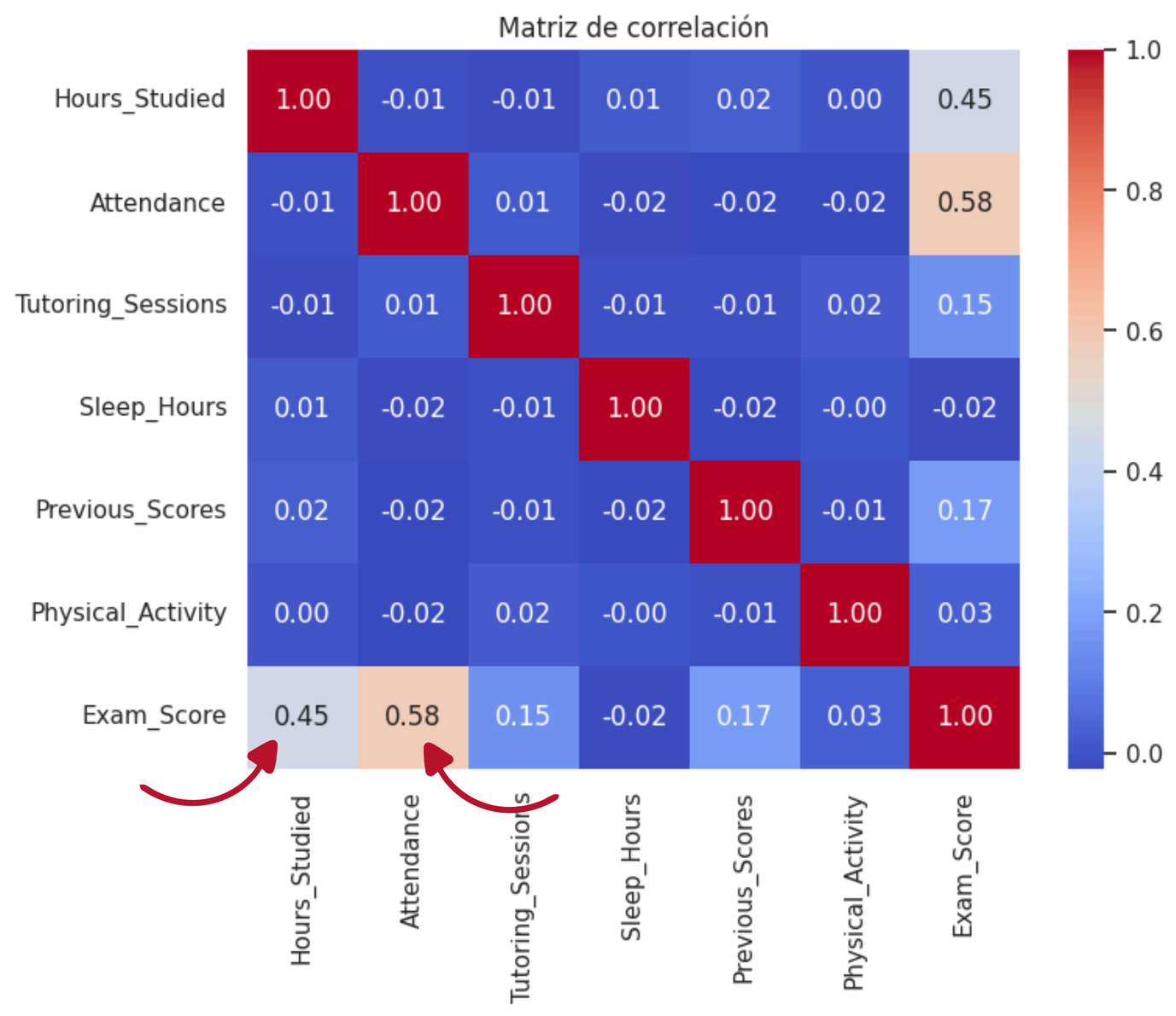
Este conjunto de datos contiene información sobre sus hábitos de estudio, asistencia, participación de los padres y otros elementos que impactan en su éxito académico.

Objetivo

Predecir el puntaje en los exámenes de los estudiantes (**Exam_Score**), identificando y comprendiendo los factores clave que influyen en su rendimiento académico. Se buscará construir un modelo que permita anticipar con precisión el desempeño de los estudiantes. Este enfoque no solo facilita la identificación de áreas de mejora, sino que también ofrece oportunidades para apoyar a los estudiantes de manera más efectiva, optimizando su camino hacia el éxito académico.

Hipótesis

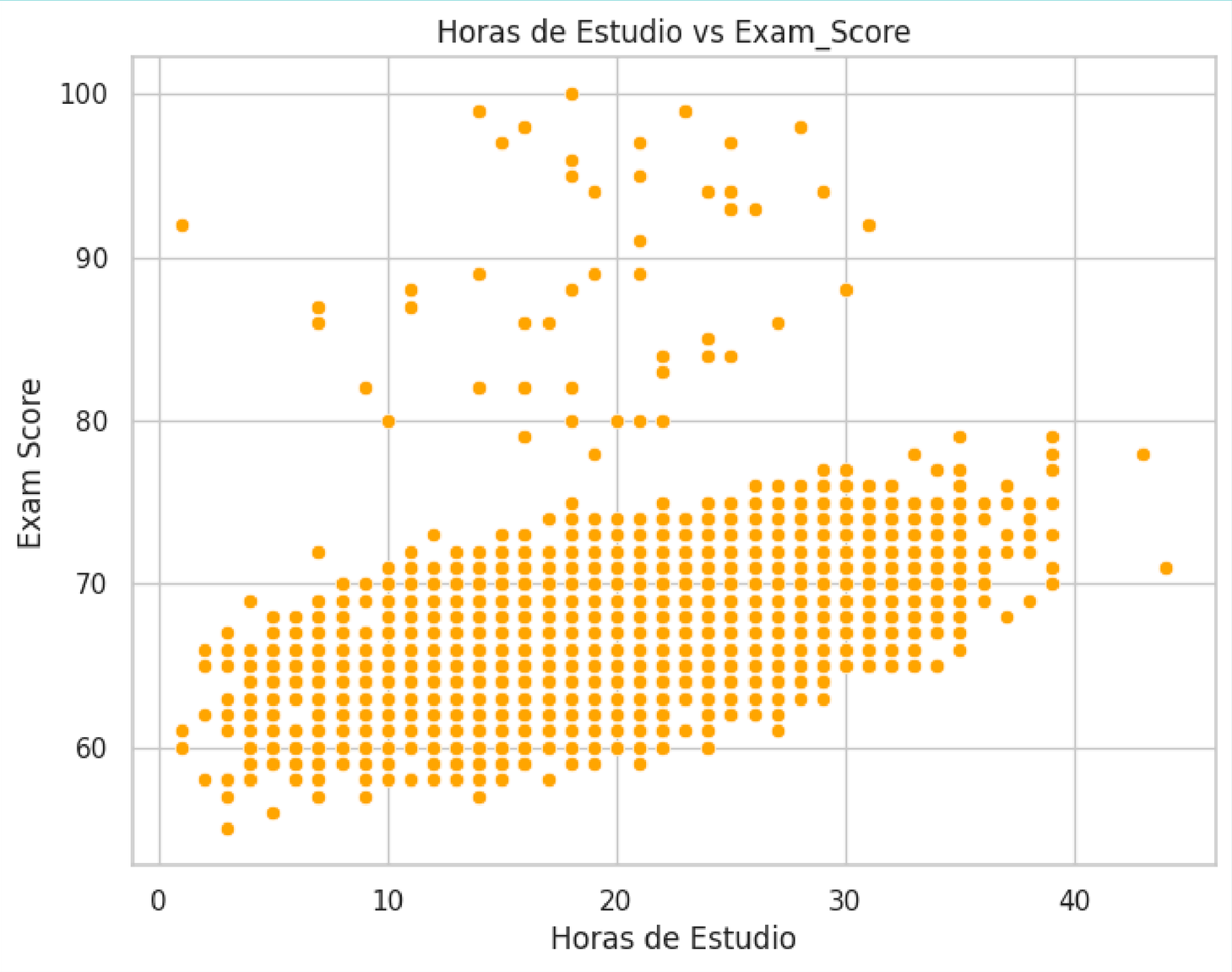
1) Los estudiantes con mayores horas de estudio tendrán un mejor desempeño en los exámenes. Se asume que cuánto mayor es la dedicación, mejores son los resultados.
2) Los estudiantes con mejor asistencia tendrán un mayor rendimiento en los exámenes. Asistir a clase es factor clave para una correcta comprensión.
3) Los estudiantes que tienen acceso a Internet y recursos educativos en línea obtendrán mejores puntajes en los exámenes. Estos recursos pueden facilitar el estudio.
4) Los estudiantes con dificultados de aprendizaje tendrán un bajo rendimiento en los exámenes.



Conociendo el dataset

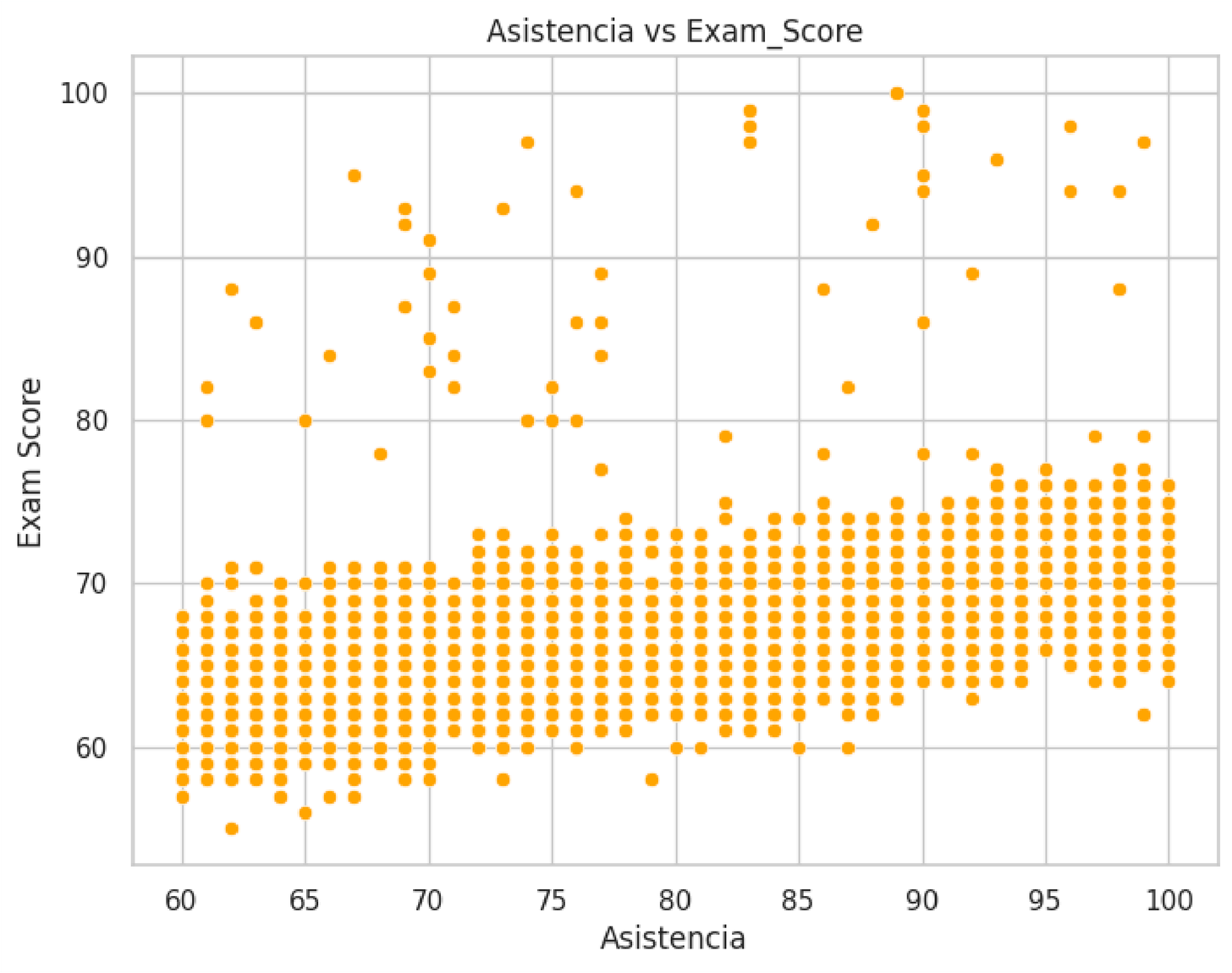
Para conocer la relación entre la variable objetivo, Exam_Score, y el resto de los factores, se estudia la matriz de correlación:

Esta matriz muestra la relación entre variables numéricas, en este caso el resultado de los exámenes presenta una relación más fuerte con las horas de estudio y la asistencia a clases.



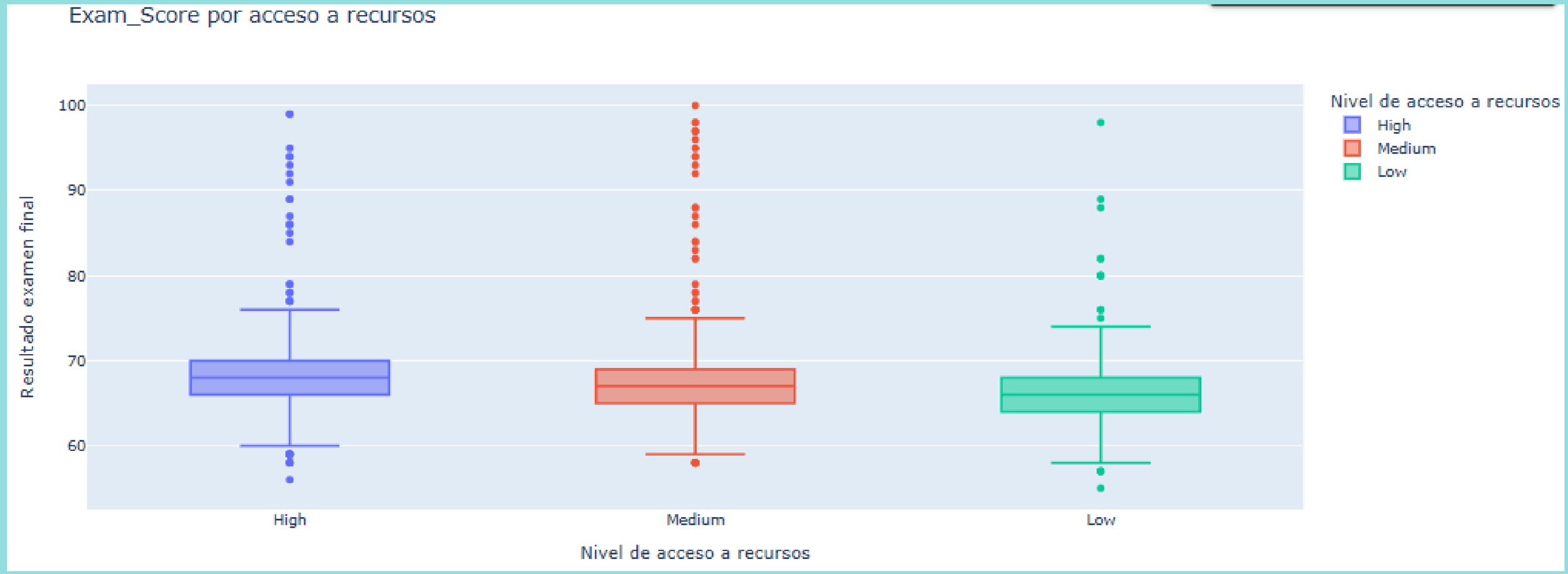
Siguiendo el resultado de la matriz, se analiza la relación entre el resultado de los exámenes y las horas de estudio:

Este scatterplot sugiere que a mayor horas de estudio se visualizan mejores resultados en los exámenes, notándose un crecimiento más significativo a partir de las 30 horas de estudio semanales. Este primer resultado apoyaría la **Hipótesis 1).**

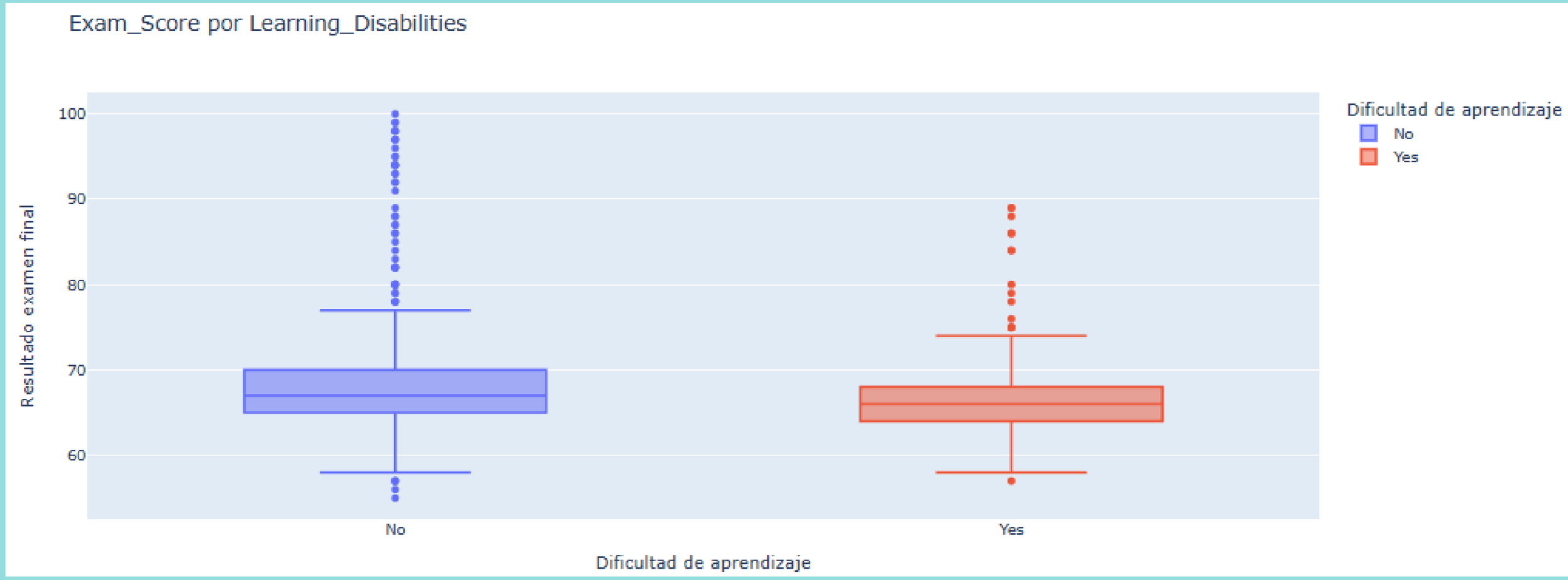


En segundo lugar, se analiza la relación entre el resultado de los exámenes y la assistencia a clase:

Este scatterplot muestra también que cuanto mayor es el porcentaje de asistencia, mejor es el resultado. Con porcentajes de asistencia mayores a 90 parece acercarse un resultado en los exámenes de 80 puntos. Este primer resultado apoyaría la **Hipótesis 2).**



El gráfico muestra cómo afecta el nivel de acceso a recursos en el resultado de los estudiantes.
La mediana de los estudiantes que tienen bajo acceso, es ligeramente menor a los que tienen alto acceso. La población de estudiantes con un nivel de acceso alto, se concentra en resultados mayores, la escasez de recursos lleva a que los estudiantes no rindan de la mejor manera en los exámenes.
Esta primera observación apoyaría a la **Hipótesis 3),** entendiendo que no se observa aún una diferencia singificativa.



Este gráfico muestra cómo las dificultades de aprendizaje se reflejan directamente en el resultado de los exámenes. Si bien esta conclusión es la esperada para la **Hipótesis 4)**, en el gráfico se puede ver como la media de los estudiantes que no presentan dificultades de aprendizaje es ligeramente mayor de aquellos que sí lo hacen.
Este es un punto de reflexión para los colegios, invertir en espacio y herramientas que disminuyan esta grieta y den oportunidades a todos los estudiantes por igual.

PRE PROCESAMIENTO:

El pre procesamiento de datos es un paso fundamental en la construcción de modelos, garantiza que los datos estén en un formato adecuado para el análisis. Este proceso incluye varios pasos:

1 OUTLIERS

Se calcula el porcentaje de outliers. Removiendo outliers no se distorsiona la distribución del objetivo Exam_Score y se conserva el 91.7% del dataset.

2 VALORES NULOS

Se calcula el porcentaje de valores nulos por columna. Solo tres de ellas presentaron valores nulos con un bajo porcentaje. Se procede a removerlos.

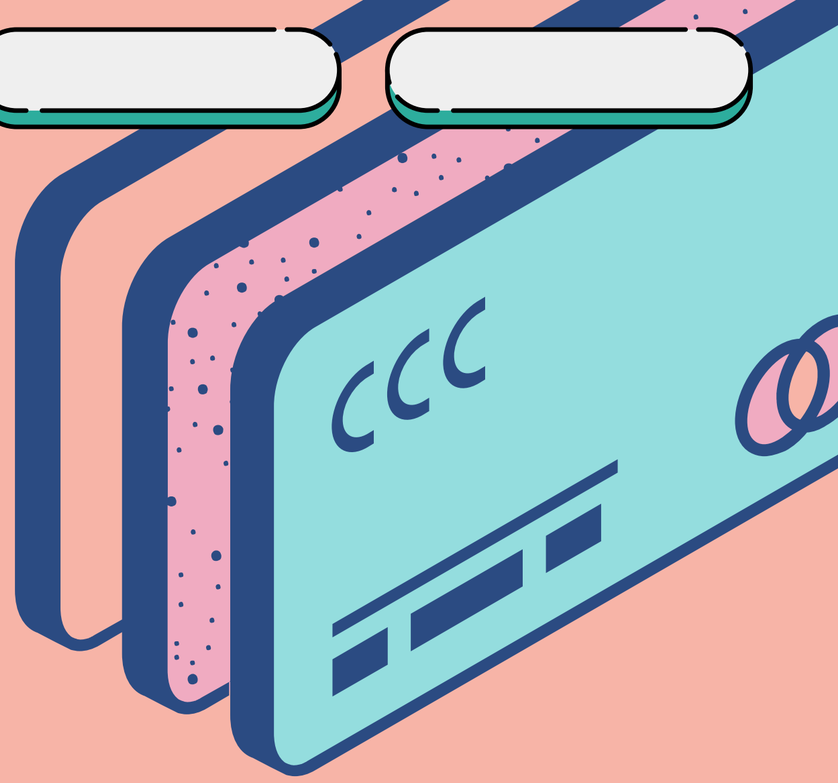
3 COLUMNAS IRRELEVANTES

Se muestran las columnas que poseen valores únicos para todos los registros. En caso de que existan, serán irrelevantes. El dataset no presenta columnas irrelevantes.

4 ENCODING Y NORMALIZACIÓN

Se utilizan para mejorar la calidad de los datos y que los modelos puedan interpretarlos. Se convirtieron datos no numéricos en un formato que los modelos puedan procesar, y se evitó que las diferencias de magnitud afecten el rendimiento de los modelos.

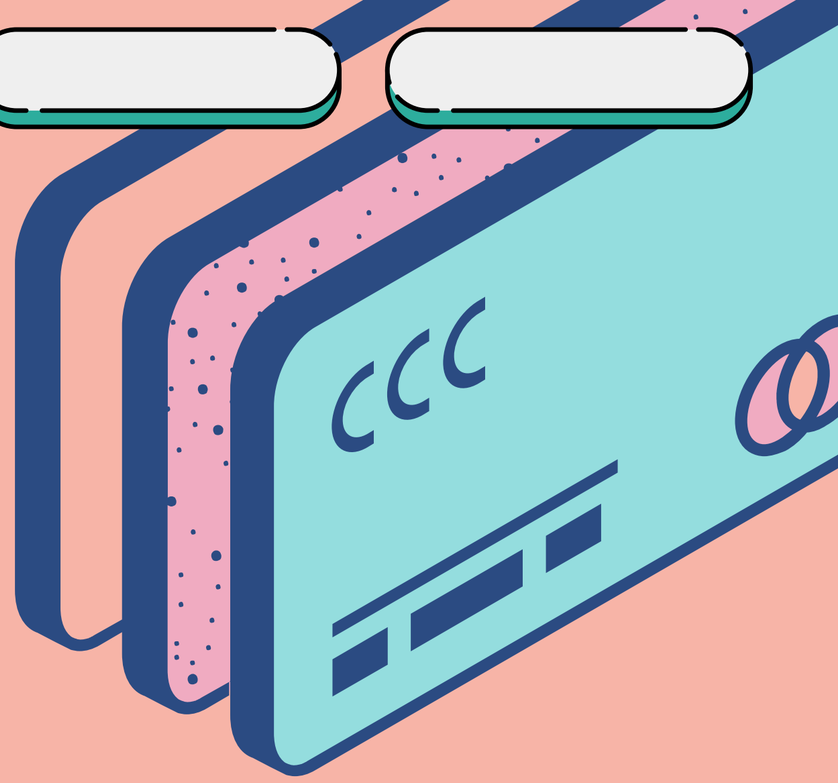
Entrenando modelos



El objetivo es predecir **Exam_Score** (numérica), se usarán modelos de regresión y se compararán las siguientes métricas: Error Cuadrático Medio (MSE), Error Absoluto Medio (MAE), R^2 (coeficiente de determinación), y el Tiempo de ejecución.
Primero se divide el dataset en entrenamiento y prueba y luego se entrenan los modelos para obtener la siguiente tabla comparativa:

Modelo	MSE	MAE	R2	Tiempo
KNN	2.345471	1.230040	0.766481	0.120518
Random Forest	1.116251	0.845454	0.888864	3.627347
Árbol de decisión	3.054826	1.325528	0.695856	0.069755
Regresión Lineal	0.100985	0.268870	0.989946	0.068321
Catboost	0.271676	0.396924	0.972951	25.400086

Comparando modelos



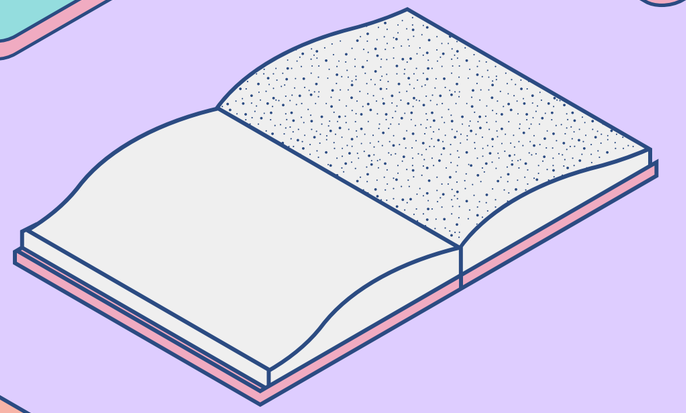
- **Regresión Lineal** es el que mejor explica los datos y tiene los errores más bajos, lo hace muy atractivo y sugiere que el modelo se ajusta bien a los datos.
- **Random Forest** es un modelo robusto y potente, su tiempo de entrenamiento es algo mayor que el de la regresión lineal.
- **CatBoost** también tiene un rendimiento con un R^2 muy alto (0.972951) y un error bajo, pero su tiempo de entrenamiento es mucho mayor en comparación con los otros modelos, lo que puede convertirse en un problema.
- **Árbol de Decisión** parece ser el modelo más débil en cuanto a precisión y capacidad para explicar la variabilidad en los datos. Es rápido pero menos adecuado para este tipo de tarea.

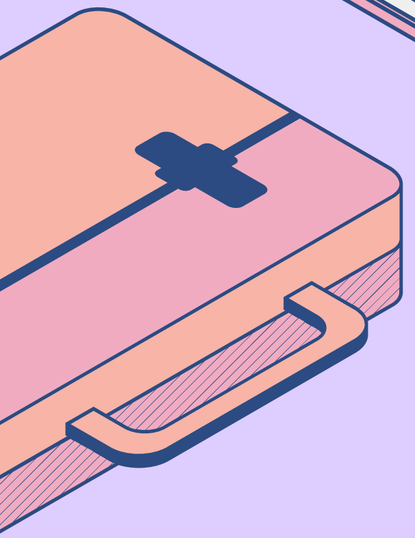
Como primer resultado, **Regresión Lineal** es un modelo rápido y de buen rendimiento pareciendo ser el mejor equilibrio

Validación cruzada

Se probará la capacidad de los modelos para predecir datos nuevos con el fin de detectar problemas como sobreajuste o sesgo y brindar una idea de cómo se comportarán con un conjunto de datos desconocido:

- Overfitting en **Random Forest** y **Árbol de Decisión**: las métricas suigeren que ambos modelos están sobreajustando los datos, con tiempo relativamente alto en Random Forest. Es probable que estos modelos hayan memorizado los datos de entrenamiento.
- **CatBoost:** tiene un buen rendimiento, pero su tiempo de entrenamiento es muy alto.
- **Regresión Lineal:** las métricas son demasiado optimistas y podría haber overfitting. Es un modelo muy eficiente y rápido.
- **KNN**: tiene un buen rendimiento, con métricas de error no tan bajas como otros modelos, el tiempo de entrenamiento es razonable.





Si el rendimiento es clave y el tiempo no es un inconveniente, CatBoost parece ser una buena opción.
Si la velocidad es un factor más importante, KNN sería más adecuado.
El resto de los modelos podrían necesitar ajustes debido al overfitting.

Como recomendación para siguientes pasos se podría continuar con un ajuste de hiperparámetros, realizando una búsqueda de hiperparámetros para optimizar el rendimiento de los modelos, en Random Forest y Árbol de Decisión.
También se poría probar con modelos adicionales, más complejos, para comparar su rendimiento y analizar si se pueden obtener mejores resultados.

¡MUCHAS GRACIAS!